Sex Equality in the United States

Boxiang Tang (ID: 1006485475)

report and codes are available at:
https://github.com/BOXIANGTANG/STA304-FINAL

20/12/2020

Contents

0.1 Abstract

(Note: In order to follow the given instructions and make this report less compelling to the readers, I have replaced all of the "I" and "my" with "we" and "our", however, this work is an independent work, not a group work.)

The United States has already put a lot of efforts into the eliminations of the sexism and the improvement of women's living standard since 1970s. However, nowdays sex discrimination is still raising concern among the American citizens. This study aims to explore the progresses that the US had made on the sex equality, and tries to identify the remaining sexism issues which the American female still suffers from. The whole study will be based on the latest US census data, but because of the equipment limitations, only 10000 randomly selected data points have been used. In this analysis, for generating stronger causal inference we not only apply the standard model regression approach, but also employ the propensity score matching regression technique. Thus, the randomness of the distribution of sex variable could be ensured, and we could see more clearly that how the difference in sex variable could cause a change in the result of our estiamted objects: probability of being employed, probability of stucking in the poverty and the expected income values. Our chosen models for this study are the logistic regression model and the mutiple linear regression model. Also, we have introduced four demopgraphic predictors: race, sex, age, and education level into those models. Afte the process of plugging in the data, we compare the results from the prviously mentioned two approaches, and finally come to the conclusion that although for eliminating the sex discriminations, the United States did make a huge progress on offering equal job opportunities for both sexes; however, the remaining sexism on the salary allocations are still holding back the improvement of American women's living standards.

0.2 Keywords

sex equality, standard modeling approach, propensity score matching, causal inference, probability of being employed, probability of stucking in the poverty, expected income

0.3 Introduction

Historically, females had been unequally treated by law and conventions for a long time. In US, since early 1970's rising women's movement has forced Supreme Court to put the sex-based equality idea in both Fourteenth Amendment and Equal Rights Amendent for eliminating the female citizens' concerns. (Law, 1984). In addition, at that time, the amendment to Title Vii of the United States Civil Rights Act of 1964 (i.e: The Pregnancy Discrimination Act of 1978) even further strengthen the protection on women. It seems that the United States has already come a long way towards achieving the sex equality. However, nowadays there are still many remaining different-look discriminations on female subgroups that postpone the women's pace to a higher level living standard. (Kramer, 2014). And those sex-equality based laws such as the Title Vii are still criticized by many for their effectiveness in defending women's rights. (Schultz, 2015). All of this makes us intentionally wonder that how is the sex equality in US today.

Traditionally, when study about sex discrimination, researchers like to start from job access or income gap. Because in the modern society, two major areas of sex discrimination are having same chance to get a certain job and gaining same wages with man for identical jobs. (Firth, 1982). Therefore for this study, we will continue to apply model analysis on probability of being employed and expected income of both sex; calculate and compare the difference between the female and male.

Nevertheless, to deeply explore the sex equality circumstances in the United States, another learning object has also been used for the study - the probability for a woman be in the poverty status. Since the recent decades, "feminisation of poverty" which indicates a rapid increase in proportion of women who are suffering from poverty at a global scale has become glaringly obvious. (Chant, 2003). This un1040equal distribution in the poverty proportion between the male and female reflects a cruel fact that although there is a rising concern about women's rights, the burden and prejudice on the female has not improved much. Hence, it is reasonable for us to add this topic in this study for further evaluating the sex equality in the US.

For this study report, it will be produced based on the census data from the IPUMS. More information about the used data will be given in the Methodology section. In addition, for profoundly investigate the link between sex and the difference in employment rate, expected income and the probability of being in poverty, both standard regression techniques and propensity score matching regression study are applied and compared. Detailed descriptions about those things will also be included in the Methodology section. Results of the two study methods' comparison and some further analysis will be provided in the Results section. Finally, a brief discussion about the results, relevant findings, study's drawbacks and potential future improvements will be covered in the Discussion section.

0.4 Methodology

0.4.1 i. Data

This report is based on the latest IPUMS data(2019), which is a census records collected by the interdisciplinary research center at the University of Minnesota. Because it is a cencus data, we may treat our used database is unbiased and representative.The only drawback for this dataset is that its sizs is too large, thus needs our equipments to have strong computing power.

However, our equipment doesnot meet that requiremnet, thus we cannot apply the original large dataset in our model (especially when doing propensity score matching). We would need to use SRS method (simple random sampling) to randomly extracted data points for our subsequent analysis. The SRS is done by simply applying the "sample()" function in R. Therefore, our population would be all American citizens,

Table 1: Table 1

	female(N=1633690)	male(N=1571522)	Overall(N=3205212)
Race			
american indian or alaska native	16437 (1.0%)	16400 (1.0%)	32837 (1.0%)
black/african american/negro	156075 (9.6%)	144832 (9.2%)	300907 (9.4%)
chinese	24627 (1.5%)	21083 (1.3%)	45710 (1.4%)
japanese	4654 (0.3%)	3374 (0.2%)	8028 (0.3%)
other asian or pacific islander	67365 (4.1%)	60977 (3.9%)	128342 (4.0%)
other race, nec	56241 (3.4%)	57562 (3.7%)	113803 (3.6%)
three or more major races	6083 (0.4%)	5766 (0.4%)	11849 (0.4%)
two major races	44218 (2.7%)	43852 (2.8%)	88070 (2.7%)
white	1257990 (77.0%)	1217676 (77.5%)	2475666 (77.2%)
Employment Status			
employed	734248 (44.9%)	798940 (50.8%)	1533188 (47.8%)
n/a	260195 (15.9%)	273226 (17.4%)	533421 (16.6%)
not in labor force	608209 (37.2%)	463918 (29.5%)	1072127 (33.4%)
unemployed	31038 (1.9%)	35438 (2.3%)	66476 (2.1%)
Education Level			
1 year of college	194169 (11.9%)	183500 (11.7%)	377669 (11.8%)
2 years of college	123401 (7.6%)	94145 (6.0%)	217546 (6.8%)
4 years of college	268906 (16.5%)	237766 (15.1%)	506672 (15.8%)
5+ years of college	169469 (10.4%)	150835 (9.6%)	320304 (10.0%)
grade 10	35140 (2.2%)	39501 (2.5%)	74641 (2.3%)
grade 11	38973 (2.4%)	43703 (2.8%)	82676 (2.6%)
grade 12	476923 (29.2%)	478336 (30.4%)	955259 (29.8%)
grade 5, 6, 7, or 8	101724 (6.2%)	106377 (6.8%)	208101 (6.5%)
grade 9	31620 (1.9%)	35405 (2.3%)	67025 (2.1%)
n/a or no schooling	78845 (4.8%)	81303 (5.2%)	160148 (5.0%)
nursery school to grade 4	114520 (7.0%)	120651 (7.7%)	235171 (7.3%)
Wage and Salary Income			
Mean (SD)	180000 (359000)	208000 (369000)	193000 (364000)
Median [Min, Max]	15000 [0, 1000000]	33000 [0, 1000000]	24000 [0, 1000000]
Poverty Status			
Mean (SD)	316 (172)	324 (173)	320 (172)
Median [Min, Max]	336 [0, 501]	356 [0, 501]	346 [0, 501]
Age			
Mean (SD)	43.6 (23.8)	41.4 (23.2)	42.5 (23.5)
Median [Min, Max]	45.0 [1.00, 96.0]	42.0 [1.00, 96.0]	43.0 [1.00, 96.0]

our sampling frame would be the original large dataset, and our sample would be those randomly selected people. The final decided simple size is 10000, simply because it is very equipment-friendly and large enough for us to generate representative results.

0.4.1.1 1. General:

Notice that the total data population are almost equally distributed between both sexes (i.e: No one has significantly more people in total than the other). In summery, the used dataset contains 3 categorical variables with multiple levels(i.e: Race, Employment Status, Educatuion Level) 1 dummy variable (i.e: sex), and 3 numerical variables (i.e: Wage and Salary Income, Poverty Status, Age).

0.4.1.2 2. Employment Status:

Recall that our first analysis object is the probability of being employed for both sexes, here because our interested variable is categorical and we want to estimate the probability of being certain category, hence the logistic model would be perpor for this estimation.

About the information provided by the table, there is an obovious difference between the proportion of two sexes getting employed (clearly the male is higher than the female), which indicates the general employment status of women is worse than men. Also, note that a large amount of females are even not in the labor force, which implies that women are less likely to participate in the labour market.

Notice that, in order to apply this variable in the logistic model to calculate the probabilty of being employed for both two sexes, we have to convert it into a binary variable first (only contains "employed" or "unemployed"). Thus the extra levels like "n/a" and "not in labour force" should be deleted before the modeling.

0.4.1.3 3. Wage and Salary Income:

Our second interested object is expected income of both sexes. Obviously, it is a numnerical variable, therefore the linearly regression model should be applicable here. By the table, we could see that there is a huge gap between women's average income and men's average income, futhermore the gap between their median income are even much larger.(Men's median income is more than twice as much as women's). Thus it would be resonable for us to suspect that there is going to be a quite big gap between two sexes' expected income.

0.4.1.4 4. Poverty Status:

Notice we got a numberical variable here, however what we want to estimate is the probability of being poverty which is categorical. Thus we should transform the data into binary form first, and then properly apply the logistic model. This variable was calculated by the formula: **(real income / poverty threshold income) * 100%** , so if its value lower than or euqal to 100, it means the corresponded person is in the poverty status. By this way, we could easily convert this numerical variable into a dummy variable.

Although both average and median poverty indicator values of the felmale are slightly lower than the male, both sexes actually performed quite similar at this part. Hence we may intuitively make the prediction that the difference in the probability of being poverty between two sexes maybe not very large.

0.4.1.5 5. Extra:

From data of "Race", we could see that in the US the population distribution of each race group for two sexes are very similar, and the majority race group in both sexes is the white.

For "Education level", women hold a higher percentage for taking advanced education (college-level). Hoever, most of people in both sexes are highly educated.

About "Age", notice that both the average age and the mdeian age of the female are higher than the male. Also, overall the average age in the US is over 40 years old which could be a potential indicator of the population aging.

0.4.2 ii. Model

0.4.2.1 1. Variable Choosing: The first crucial step for modeling is choosing appropriate variables. Recall that our estimated objects (response variables) are probability of being employed, expected income and probability of stucking in poverty, and we would focus on the difference of those values between two sexes. In order to calculate and compare the difference, undoubtedly we have to include sex variable. In addition, nowdays, age discrimination is not only manifested in more willingness to recruit younger applicants, but also in forced order workers to retire or demote them to less remunerative positions (Johnson & Neumark, 1999). Thus, the order people are more likely to have lower values in our three estimated objects, which indicate a potential negative correlation between them, hence we need to count the "age" in. Also, several studies have reported that in the United States, lots of people from the racial minorities had personally been passed over for a job or promotion simply because of their ethnicity (Pager & Shepherd, 2008), which implies possible correlation between race and those three response variables. Furthermore, recent study ILOSTAT (International Labour Organization) shows that people with advanced education levels are more likely to be employed and get adequate earnings than those who only have a basic education level or less (Gammarano, 2020). Therefore we are reasonable to believe that the education level is strongly correlated with our three learning objects, thus it also should be included in the model. All in all, our selected predictors would be: sex, age, race, and education level.

0.4.2.2 2. Modeling:

0.4.2.2.1 a. Overall: Other than using an ordinary approach that simply setting up the regression models (linear & logistic); for this study, we will also do it in a different way which is first applying "propensity score matching" to organize the data, and then using the organized data to re[eat the previous standard method. By wielding those two methods and comparing their results, we will not only get a stronger causal inference between the sex difference and the expected value differences of our three estimated objects, but also examine whether sexes are distributed randomly or not. Philosophically we usually assume that sexes are distributed randomly, however, this may not be the case. In order to use sex variable as a random "treatment" to generate a valid causality result, we have to make sure the randomness of the sexes in our data. Note that, because our chosen data is representative, thus our conclusion is possible to be extended to the population level. More specific information will be elaborated in the following subsections.

0.4.2.2.2 b. Logistic Regression Modeling: In statistics, logistic regression modeling is often used to model the binary outcome variables and estimate the probability of non- based level (binary value equal to "1") event occurring. Previously we have mentioned that we want to estimate the probability of being employed and the probability of stucking in poverty for both sexes, and in order to do that, we have already transformed " employment status" and "poverty status" into dichotomous format. However, for predicting our target probabilities, we need a further step to set the level "employed" in "employment status" variable, and level "in poverty" in "poverty status" variable to be non-based level. After that, we could properly apply them in logistic models respectively by the formula:

$log(\frac{p}{1-p}) = \hat{\beta_0}+\hat{\beta_a}\ x_a+\hat{\beta_{r.b}}x_{r.b}+\hat{\beta_{r.c}}x_{r.c}+....+\hat{\beta_{r.w}}x_{r.w}+\hat{\beta_{e.2}}x_{e.2}+\hat{\beta_{e.4}}x_{e.4}+...+\hat{\beta_{e.n2}}x_{e.n2}+\hat{\beta_{s.m}}x_{s.m}$

$log(\frac{p}{1-p})$: log odds of being "employed" (or "in poverty")

$\hat{\beta_0}$: log odds of being "employed" (or "in poverty") for a female american indian or alaska native whose education level is 1 year of college and age is 0. (not practical)

$\hat{\beta_{r.i}}$: Holding all the other predictors fixed, the expected value of increase in the log odds when level "i" in predictor "race" is satified (i.e: level"i" is True : $x_{r.i} = 1$; note here "i" could be "black/african american/negro","chinese"...)

$\hat{\beta_{e.i}}$: Holding all the other predictors fixed, the expected value of increase in the log odds when level "i" in predictor "education level" is satified (i.e: level"i" is True : $x_{e.i} = 1$; note here "i" could be "2 years of college","4 years of college"...)

$\widehat{\beta_{s.i}}$: Holding all the other predictors fixed, the expected value of increase in the log odds when level "i" in predictor "sex" is satified (i.e: level"i" is True : $x_{e.i} = 1$; note here "i" could be "male")

$\widehat{\beta_a}$: Holding all the other predictors fixed, when age increase by 1, the expected value of increase in the log odds.

x_a: predictor "age"

$x_{r.i}$: level "i" of predictor "race" (level "i" is True: $x_{r.i} = 1$, level "i" is False: $x_{r.i} = 0$)

$x_{e.i}$: level "i" of predictor "education level" (level "i" is True: $x_{e.i} = 1$, level "i" is False: $x_{e.i} = 0$)

$x_{s.i}$: level "i" of predictor "sex" (level "i" is True: $x_{s.i} = 1$, level "i" is False: $x_{s.i} = 0$)

Dummy Variable Table:

Level (race)	$x_{r.b}$	$x_{r.c}$	$x_{r.j}$	$x_{r.o1}$	$x_{r.o2}$	$x_{r.t1}$	$x_{r.t2}$	$x_{r.w}$
american indian or alaska native	0	0	0	0	0	0	0	0
black/african american/ negro	1	0	0	0	0	0	0	0
chinese	0	1	0	0	0	0	0	0
japanese	0	0	1	0	0	0	0	0
other asian or pacific islander	0	0	0	1	0	0	0	0
other race, nec	0	0	0	0	1	0	0	0
three or more major races	0	0	0	0	0	1	0	0
two major races	0	0	0	0	0	0	1	0
white	0	0	0	0	0	0	0	1

Level(education level)	$x_{e.2}$	$x_{e.4}$	$x_{e.5}$	$x_{e.g10}$	$x_{r.g11}$	$x_{e.g12}$	$x_{e.g8}$	$x_{e.g9}$	$x_{e.n1}$	$x_{e.n2}$
1 year of college	0	0	0	0	0	0	0	0	0	0
2 years of college	1	0	0	0	0	0	0	0	0	0
4 years of college	0	1	0	0	0	0	0	0	0	0
5+ years of college	0	0	1	0	0	0	0	0	0	0
grade 10	0	0	0	1	0	0	0	0	0	0
grade 11	0	0	0	0	1	0	0	0	0	0
grade 12	0	0	0	0	0	1	0	0	0	0
grade 5, 6, 7, or 8	0	0	0	0	0	0	1	0	0	0
grade 9	0	0	0	0	0	0	0	1	0	0
n/a or no schooling	0	0	0	0	0	0	0	0	1	0
nursery school to grade 4	0	0	0	0	0	0	0	0	0	1

Levels (sex)	$x_{s.m}$
male	1
female	0

But be careful that by this formula,we would only get the log odds of our target outcomes, hence for generating the wanted probabilities, we should apply some mathematical methods to transform the formula to :

$p = \frac{e^{\hat{\beta_0}+\hat{\beta_a}x_a+\dots}}{1-e^{\hat{\beta_0}+\hat{\beta_a}x_a+\dots}}$

and calculate the values of "p" for both "employed" and "in poverty" respectively.

0.4.2.2.3 c. Multiple Linear Regression Modeling: Multiple linear regression is a statistical technique that uses several predictors (explanatory variables) to predict the outcome of a numeric response variable. Recall that one of our three estimated objects - "the expected income" is a numerical variable. Hence we could properly apply it in this model as a response variable by formula:

$\hat{Y_i} = \hat{\beta_0}+\hat{\beta_a}\ x_a+\hat{\beta_{r.b}}x_{r.b}+\hat{\beta_{r.c}}x_{r.c}+\dots.+\hat{\beta_{r.w}}x_{r.w}+\hat{\beta_{e.2}}x_{e.2}+\hat{\beta_{e.4}}x_{e.4}+\dots+\hat{\beta_{e.n2}}x_{e.n2}+\hat{\beta_{s.m}}x_{s.m}$

$\hat{Y_i}$: the epected income value by given predictors

(Note: the interpretations of the other factors are similiar to the logistic model, except that the meaning of predictors' coefficients has changed to be the "increase in the expected income" not log odds anymore; also the Dummy variable tables here would be same as the previous ones.)

0.4.2.2.4 d. Propensity Score Matching: Generally speaking, propensity score matching is a statistical technique that matches treated and controlled observations on the estimated probability of being treated which is called "propensity score". It is a quite popular approach to estimate causal treatment effects. To be specific, the method contains two key concepts: matching and propensity score. Matching is a data organizing method which is selecting the subset data of treated and control groups that are similar to one another in the characteristics excluding the actual treatment (Stuart & Rubin & Osborne, 2007). It is a common approach to generate causal inference by using the data which does not contain random assignment of treatments to the units (i.e observational data) (Rubin, 1973). Propensity score is the probability of receiving a treatment conditional on the observed covariates (i.e the observed characteristics except the actual treatment) (Lee & Lessler & Stuart, 2010). Usually we use the logistic regression to calculate it by a rough formula:

$Treatment = logit^{-1}(O.C_1+O.C_2+\dots)$

$O.C_i$: Observed covariate "i" (e.g "age", "employment status" ...)

For our study, because we mainly focus on the differences caused by change of sex variable, thus our treatment would be sex (i.e treat non-based level "Male" as "treated", base level "Female" as " not treated"), then the observed covariates would be the rest of our chosen variables:

$Sex = logit^{-1}(Age+EducationLevel+Race+EmployedStatus+PovertyStatus+Income)$

Mathematically speaking, it should be:

$log(\frac{p}{1-p}) = \hat{\beta_0}+\hat{\beta_a}\ x_a+\hat{\beta_{r.b}}x_{r.b}+\hat{\beta_{r.c}}x_{r.c}+\dots.+\hat{\beta_{r.w}}x_{r.w}+\hat{\beta_{e.2}}x_{e.2}+\hat{\beta_{e.4}}x_{e.4}+\dots+\hat{\beta_{e.n2}}x_{e.n2}+\hat{\beta_{e.e}}x_{e.e}+\hat{\beta_{p.s}}x_{i.p}+\hat{\beta_{in}}x_{in}$

$log(\frac{p}{1-p})$: log odds of being male (no practical meaning)

p: propensity score

$\hat{\beta_{e.e}}$: Holding all the other predictors fixed, the expected value of increase in the log odds when the level "employed" in predictor "employment status" is satified (i.e: level"employed" is True : $x_{e.e} = 1$)

$\hat{\beta_{p.s}}$: Holding all the other predictors fixed, the expected value of increase in the log odds when the level "in poverty" in predictor "poverty status" is satified (i.e: level"in poverty" is True : $x_{i.p} = 1$)

$\hat{\beta_{in}}$: Holding all the other predictors fixed, the expected value of increase in the log odds when there is a unit change in income

$x_{e.e}$: level "employed" of predictor "employment status" (level "employed" is True: $x_{e.e} = 1$, level "employed" is False: $x_{e.e} = 0$)

$x_{i.p}$: level "in poverty" of predictor "poverty status" (level "in poverty" is True: $x_{i.p} = 1$, level "in poverty" is False: $x_{i.p} = 0$)

x_{in}: predictor "income"

(Note: the interpretations of all the other factors are similiar to the previous logistic regression model)

Dummy Variable Tables:

Levels (employment status)	$x_{e.e}$
employed	1
unemployed	0

Levels (poverty status)	$x_{i.p}$
in poverty	1
poverty	0

(Note: all the other dummy variable tables are the same as the previous ones)

Same way to generate the final value of the propensity score p as previous mentioned in the logistic regression modeling part. Note that sex is not a normal treatment that could be contolled or not. Our purpose here is to study whether sexism exists in real life from a statistical perspective. Therefore some values we calculated such as propensity score here does not have any practical meanings, we just use it as a statistical tool for organizing data.

After the propensity score calculations for each data point, we will match the data from the treated group ("Male" group) to the data from the controlled group ("Female" group) by comparing if their propensity scores are the closest to each other. Finally we would reduce our dataset by only keeping the data points that are properly matched. This would be the end of the whole propensity score matching approach, for the next step, we will use the organized new dataset to repeat the ordinary modeling method again to see whether the results are consistent with those obtained by previous technique (directly applying ordinary modeling method). Note, all the previous mentioned processes would be done by running software: R.

Notice that, because of the equipment limitations, in order to run the propensity sore matching codes successfully, we have to reduce the dataset to a equipment-friendly size. Therefore, all of the results that appear in the Result section are calculated based on 10000 randomly selected data points which comes from the original large data set.

0.4.2.3 3. Evaluation of Models & Potential Aternative Modeling Methods:

Models	AIC Values
sd method logi(employed)	2276.3
ps method logi(employed)	2193.5
sd method logi(poverty)	3671
ps method logi(poverty)	3646.6

Models	Adjusted R square Values
sd method line(income)	0.1969
ps method line(income)	0.212

Both the Akaike information criterion (AIC) values and adjusted R square values are mathmetical indicators for evaluating how well the model fits the given data. To make the model more proper for a given dataset we would always prefer lower AIC values or higher adjusted R square values. By the previous tables, we could see that the data which has been matched by propensity score makes our chosen models perform even better. (i.e the AIC values of logistic models have decreased, and the adjusted R square values of linear model have increaded)

Except the ordinary logistic or linear models, actually we could also apply Baysian regression models for this study. By this approach we could use our owned knowledge and information about the data to provid it a proper prior. Also, this approach gives us inferences which are conditional on the data and are exact. In addition, it not only obeys the likelihood principle, but also provides interpretable results. However, in order to apply Baysian approach properly, we have to select a appropriate prior which requires us to have a good understanding of the dataset, but usually we do not have that much information about the data we wish to use. That would be the most difficult point for us to use it in this study.

0.4.2.4 4. Model Diagnostic: a. VIF Table for Linear Model (Income):

Predictors	VIF Value (GVIF)
Sex	1.011477
Education Level	1.117593
Race	1.087658
Age	1.035365

b. VIF Table for Logistic Model (Employed):

Predictors	VIF Value (GVIF)
Sex	1.013649
Education Level	1.082287
Race	1.152270
Age	1.070049

c. VIF Table for Logistic Model (Employed):

Predictors	VIF Value (GVIF)
Sex	1.015903
Education Level	1.093321
Race	1.175989
Age	1.071938

Multicollinearity is the problem of fitting models when two or more explanatory variables are highly correlated. It usually causes a high variance in the β, and thus the parameters would become insignificant due to the high variance. Also, it could make our fitted equation become unstable and cause a wrong sign for the regression coefficients.

Because we will mainly focus on p.s method results, thus here we only check the VIF value of p.s method's models. Note that, the VIF values of the models' predictors are all lower than the commen threshold value (i.e commen threslhold :5), which implies that our chosen predictors are not highly correlated with each other. Hence, our models may not have such problems. Therefore, probably they are all useful for our later analysis.

0.5 Results

0.5.1 i. Tables:

Models	Coefficients	Log Odds	Odds(net increase)	Odds (net increase percentage form)
sd method logi(employed)	$\widehat{\beta_{s.m}}$	-0.161994	-0.149553	-14.96%
ps method logi(employed)	$\widehat{\beta_{s.m}}$	-0.167207	-0.153976	-15.40%
sd method logi(poverty)	$\widehat{\beta_{s.m}}$	-0.547070	-0.421357	-42.14%
ps method logi(poverty)	$\widehat{\beta_{s.m}}$	-0.488037	-0.386170	-38.62%

0.5.1.1 a. Comment 1: Here for calculating net increase in odds, we have applied formula:

Odds Ratio = exp(Logistic Model Coefficient Values)

Percent Change in the Odds = (Odds Rstio - 1) x 100%

Notice that, although the resulted values of those two methods are slightly different, they all provide us similiar information about sex equality. For the probability of being employed, we could see that both method shout out under same race, educational level and age condition, a male is about 15% less likely to get a job than a female, which indicates that in the United States sex discriminations in the recruitment has been eliminate a lot, job opportnities are fairly evenly distributed between men and women.

However, when come to the probabitity of stucking in the poverty, the sex inequality does seem to be quite serious. Note, the results of both methods show that when the orther characteristics holdind fixed, the male would be nearly 40% less likely to fall in the poverty than th female, which implies that the women in the US are much more easily to suffer from the poverty. Hence we could say that the prorperty distribution between two sexes in America is very unequal.

Models	Coefficients	Coeficient values
sd method line(income)	$\widehat{\beta_{s.m}}$	26423.87
ps method line(income)	$\widehat{\beta_{s.m}}$	9723.32

0.5.1.2 b.Comment 2: Unlike the results from logistic regression model, there is a quite large gap occuring between the results of two methods' linear models. This difference maybe caused by unbalanced distribution of sex variables in the original data frame. For a much stronger causal inference, we would focus on the p.s. method result. Notice, the result shows that when all the other identities are same, the expected income of a male would be about 10000 dollars higher than a female, which indicates that the

income ineuqality between two sexes seems to be quite serious in the United States. This could also be one of the reasons for the "feminisation of poverty".

0.5.2 ii. Graphs:

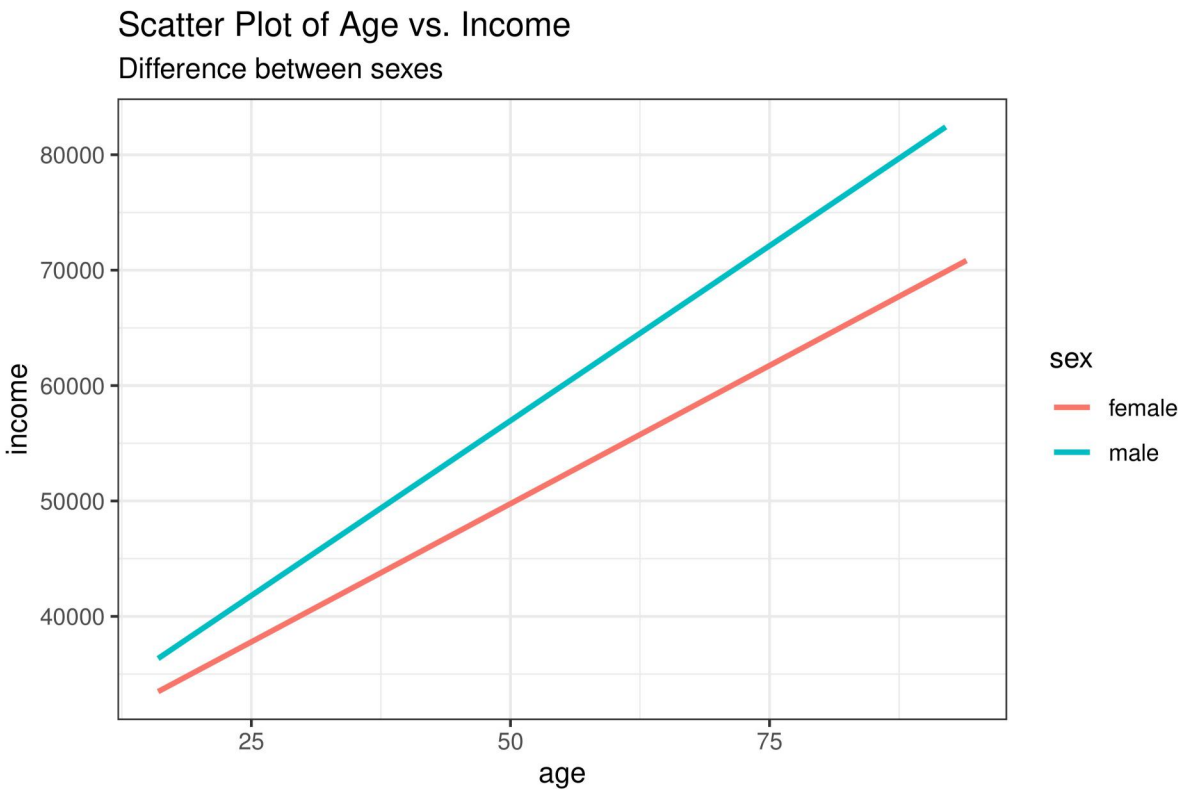

Figure 1

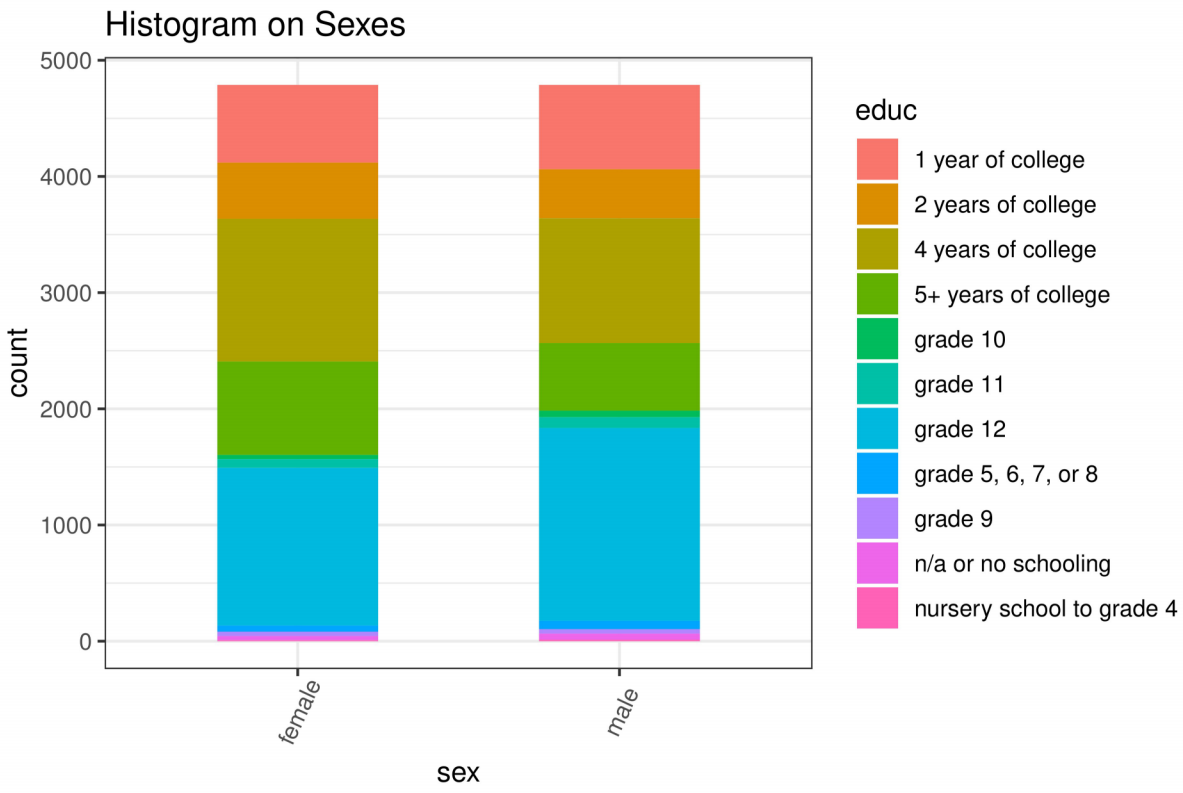
Histogram on Sexes
5000
4000
3000
2000
1000
0
count
female
male
sex
educ
1 year of college
2 years of college
4 years of college
5+ years of college
grade 10
grade 11
grade 12
grade 5, 6, 7, or 8
grade 9
n/a or no schooling
nursery school to grade 4

Figure 2

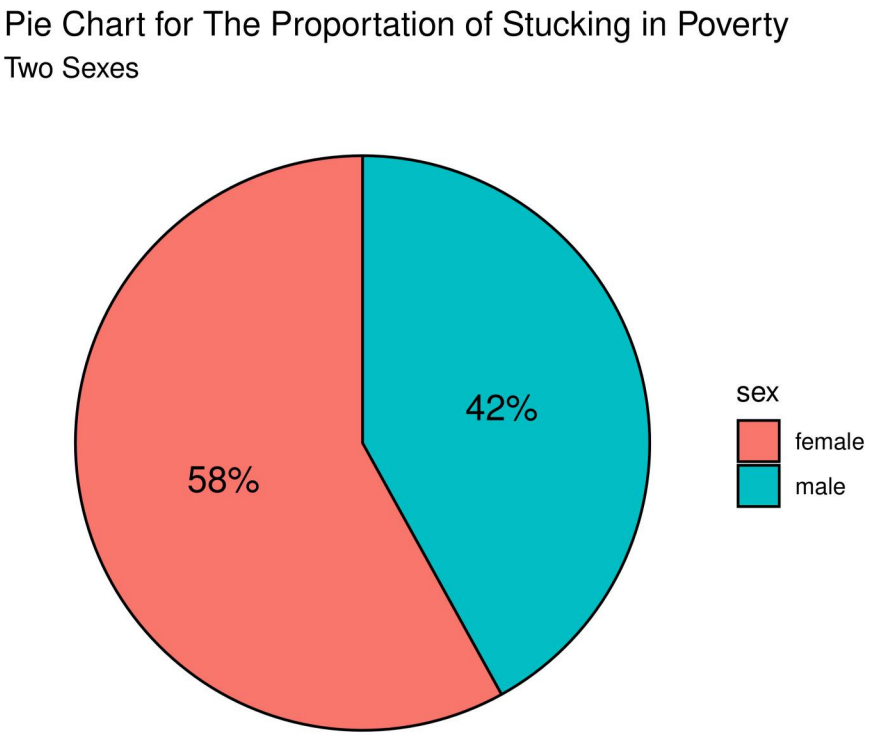
Pie Chart for The Proportation of Stucking in Poverty
Two Sexes
42%
58%
sex
female
male

Figure 3

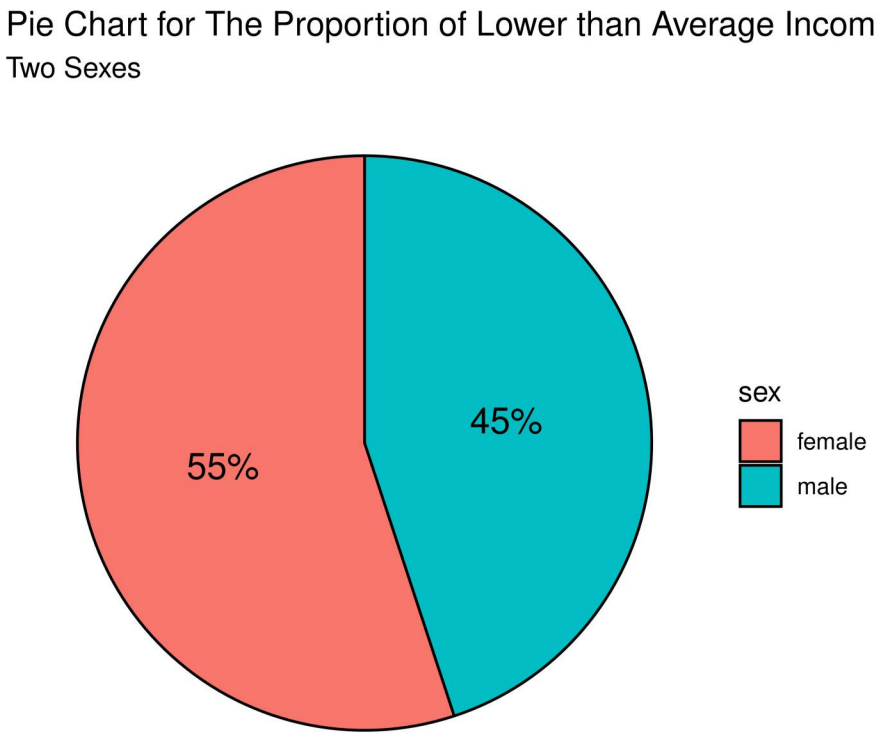
Pie Chart for The Proportion of Lower than Average Income Values
Two Sexes
55%
45%
sex
female
male

Figure 4

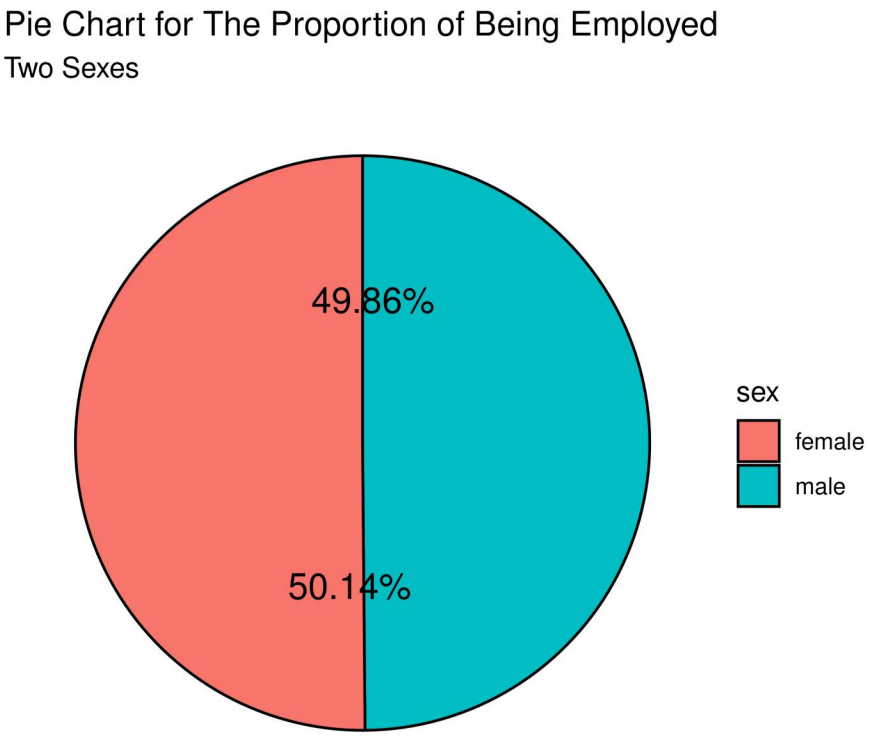

Figure 5

0.5.2.1 a. Comment 1: Notice that for more clearly observing the patterns from the best fitted lines of both sexes, we have hid all the data points in Figure 1. Approximately these two lines may have the same intercepts, which means the male and the female are probably have the same starting salary. However, obviously the men's line is steper than the women's, which indicates that as age increasing by 1, the male would get more additional salary than the female does. This result is consist with our previous statements that in the United States, nowdays the sexism is more likely to happen on the recruited women by unfair salary allocation.

0.5.2.2 b. Comment 2: By Figure 2, clearly women have a higher proportion of attending advanced education (college degrees) than men. Recent study has already shown that there is a srong positive correlation between the education level and the earning (Wolla & Sullivan, 2017). But if we combine the Figure 2 with Figure 1, it is quite strange that most of the women have very high education levels, however, their earnings are consistently lower than the men. This result further strengthen our view that there is a serious sex discrimination in the distribution of wages in the US.

0.5.2.3 c. Coomment 3: Figure 3, Figure 4 and Figure 5 make our previous view further strengthening. Note, although there are nearly same number of men and women being recruited, women are still more likely to get salaries which are lower than the average level (i.e **Average income = Total income / Total num of people**) , also they are more vulnerable to poverty than men. Hence we may say that the sex discrimination on the salary allocations is still quite serious that is currently holding back improvement of American women's living standards, albeit the distribution of jobs has already been fair enough.

Table 14: Summary of Used Data

	female(N=4788)	male(N=4788)	Overall(N=9576)
Race			
american indian or alaska native	35 (0.7%)	31 (0.6%)	66 (0.7%)
black/african american/negro	477 (10.0%)	415 (8.7%)	892 (9.3%)
chinese	67 (1.4%)	62 (1.3%)	129 (1.3%)
japanese	8 (0.2%)	9 (0.2%)	17 (0.2%)
other asian or pacific islander	200 (4.2%)	209 (4.4%)	409 (4.3%)
other race, nec	184 (3.8%)	186 (3.9%)	370 (3.9%)
three or more major races	12 (0.3%)	8 (0.2%)	20 (0.2%)
two major races	109 (2.3%)	103 (2.2%)	212 (2.2%)
white	3696 (77.2%)	3765 (78.6%)	7461 (77.9%)
Employment Status			
employed	4681 (97.8%)	4655 (97.2%)	9336 (97.5%)
n/a	0 (0%)	0 (0%)	0 (0%)
not in labor force	0 (0%)	0 (0%)	0 (0%)
unemployed	107 (2.2%)	133 (2.8%)	240 (2.5%)
Education Level			
1 year of college	669 (14.0%)	726 (15.2%)	1395 (14.6%)
2 years of college	483 (10.1%)	422 (8.8%)	905 (9.5%)
4 years of college	1229 (25.7%)	1074 (22.4%)	2303 (24.0%)
5+ years of college	804 (16.8%)	582 (12.2%)	1386 (14.5%)
grade 10	36 (0.8%)	56 (1.2%)	92 (1.0%)
grade 11	75 (1.6%)	93 (1.9%)	168 (1.8%)
grade 12	1357 (28.3%)	1659 (34.6%)	3016 (31.5%)
grade 5, 6, 7, or 8	53 (1.1%)	72 (1.5%)	125 (1.3%)
grade 9	38 (0.8%)	39 (0.8%)	77 (0.8%)
n/a or no schooling	35 (0.7%)	52 (1.1%)	87 (0.9%)
nursery school to grade 4	9 (0.2%)	13 (0.3%)	22 (0.2%)
Wage and Salary Income			
Mean (SD)	46600 (50800)	52800 (38300)	49700 (45100)
Median [Min, Max]	35000 [4.00, 571000]	45000 [30.0, 250000]	40000 [4.00, 571000]
Poverty			
Mean (SD)	369 (148)	375 (140)	372 (144)
Median [Min, Max]	425 [1.00, 501]	422 [1.00, 501]	423 [1.00, 501]
Age			
Mean (SD)	43.5 (14.8)	43.1 (15.0)	43.3 (14.9)
Median [Min, Max]	43.0 [16.0, 94.0]	43.0 [16.0, 92.0]	43.0 [16.0, 94.0]
Poverty_Status			
normal	4492 (93.8%)	4574 (95.5%)	9066 (94.7%)
poverty	296 (6.2%)	214 (4.5%)	510 (5.3%)
Propensity_Scores			
Mean (SD)	0.488 (0.106)	0.527 (0.0957)	0.507 (0.103)
Median [Min, Max]	0.488 [0.189, 0.995]	0.537 [0.178, 0.713]	0.514 [0.178, 0.995]

0.6 Discussion

0.6.1 i. Summary:

This analysis aims to study about the current situation of sex equality in the United States. To yield the result as representative as possible, the whole study has been based on the latest 2019 IPUMS cencus data. For generating a stronger causal inference, we not only apply the standard model regression approach, but also perform a propensity score matching regression technique by using sex as a "treatment". By this way, the randomness of the distribution of sex variable could be ensured, also we could see more clearly that how the difference in sex variable could cause a change in the result of the probability of being employed, the probability of stucking in poverty and the expected income values. Both the logistic regression model and mutiple linear regression model are used to estiamte those three response variables. The chosen predictors of this analysis are race, age and the education levels. Furthermore, ideally all of those processes should be done by using the orginal large data set, however, because of the equipment limitation, for simplicity, the whole study actually is only based on 10000 data points which are randomly selected from the original dataset.

0.6.2 ii. Summary of Used Data:

Here, the table shows a brief summary of our used dataset in the whole study. Recall that it is randomly select from our original census data (i.e IPUMS 2019 Census Data), and its size is 10000. Because it is randomly extracted from the census data (i.e by R function "sample()")and its size is quite large, we may reasonably treat it as an unbiased and representative dataset. However, compare to the size of the original dataset, it is relatively much smaller, thus it could be less representative than the larger one.

Compare with the previous data table, for categorical variables, the percentages of each level all have some changes, but the relative sizes are approximately consistent. Nevertheless, the "employment status" variable has a very obvious change, note here both "n/a" and "not in labour force" levels become 0, this is caused by our previous data cleaning processes. For the numerical variable, similiarly the values of mean and median are all different from before, but our previous mentioned relationships between those values of two sexes still holds. Notice, there are two extra variables in this smaller dataset - "Poverty_Status" and "Propensity_Scores" (note that the "Poverty Status" in the previous table has been changed to "Poverty" here). Those were added to the dataset in the modeling section. "Poverty_Status" shows the proportion of people who are suffering from the poverty and those who are not for both sexes. Clearly, we could see that women are more likely to be in poverty than men. Previously when we generating causal inference, we have used "sex" variable as a treatment, hence the "Propensity_Scores" here actually doesnot have any practical meaning, it is just an indicator value of showing how well two data points match.1

Generally speaking, although there are some differences between our used dataset and the orginal large datatset, most of the important data information such as relative size or correlations is still consistent. Also, becuase we get those data by SRS (Simple Random Sampling) process, in theory this dataset could be considered representative and unbiased. (However, note that actually bias could occur because of the data cleaning) Therefore, our results based on this dataset could be considered representative as well, thus the strengths of our conclusions could also be ensured.

0.6.3 iii. Conlusion:

0.6.3.1 Information Tables from Previous Pie Chart:

Sexes	Proportion of Being Employed
Male	49.86%
Female	50.14%

Sexes	Proportion of Income Lower than The Average Values
Male	45%
Female	55%

Sexes	Proportion of Stucking in The Poverty
Male	42%
Female	58%

0.6.3.2 Information Tables from Previous Scatter Plot:

Slope Difference	Relatived P-Value
144.41	$2.83 * 10^{-13}$

Recall that, in the Result section, we have concluded that although both the male and the female hold the same probability of being employed, because of the unfair wage distribution after the recruitments, women are still much more likely to stuck into the poverty than men. The provided tables and graphs even further strengthed the previous statements. Notice here, in order to prove that there is a gap between men's and women's salaries, we have built a linear model which contains an interaction term with sex variables, the meaning of this term is that the relationship between income and age are dependent on the sex variable. The results show that our previous assumption about the slope difference does hold, and that difference is very statistically significant (very small p-value). Thus, we further prove the existence of the sexism on the wages ditribution, and the that wage gap cause by sex difference are approximatly 144 dollars for each unit age increase.

Therefore, we may say that for eliminating the sex discriminations, the United States did make a huge progress on offering equal job opportunities for both sexes; however, the remaining sexism on the salary allocations are still holding back the improvement of American women's living standards. This conclusion may also be applicable to the other countries, in addtion to only focus on emphasising the equality of work chances between two sexes, maybe we should also pay close attention to limit the "sex wage gap". Thus not only equal chance for getting work, but also equal pay for the equal work.

0.6.4 iv. Weakness:

Nevertheless, there are a couple of weaknesses in this study. Firstly, as prviously mentioned, for simplicity, we only partially use the original unbiased dataset, and the size of our used data points are much smaller compare to the original one. Thus the final results may not be representative which means our conclusions may considerably differ from the reality. Secondly, about the estimate object: the probability of being employed, probably we need to calculate this ratio based on the employment status of an identical job, that would be more consistent with the "equal working opportunity" topic. But the chosen data was lack in the information of job identities, therefore maybe our analysis on that topic is biased. In addition, notice that in the "model diagnostic" part, we have only checked the multicollinearity, however, there are so many dimensions for checking the validities of the model, and testing on only one aspect is far from enough. It is possible that our models will fail the other tests, hence our inferences based on those model could be invaild as well. Finally, most of our predictors are categorical, thus the analysis (espcially th graph analysis) we can do is quite limited, which would cause our study to be less profound.

0.6.5 v. Next Steps:

For further improving this analysis, first we could use a computer with mcuh stronger computation powers and repeat the study again. By this way the equipment limitation will be solved, therefore we could do the analysis based on the original large dataset, and make our resluts much more representative. Second, to make our calculation more consistent with the study topic, we may also need to find another data frame that not only contains the information of our estiamte objects and chosen predictors, but also includs the certain job's identities, and then we could repeat our analysis again based on that dataset. Furthermore, more dimensions will be covered when diagnosing models. For linear models we could check whether the Gauss-Markov assumptions hold or not by using diagnostic plots, for logistic model we could reform the models and check its AIC OR BIC values to find the best fitted models for the given dataset. Last but not the least, more reasonable numerical predictors should be added into the models for drawing more valuable graphs, thus based on those extra provided assets we could significantly increase the deepness of our study.

0.7 References

Law, S.A.(1984). Rethinking sex and the constitution. Penn Law Journals, 132(5), 955-1040.

Kramer, Z.A(2014). The new sex discirimination. Duke Law Journal, 63(4), 891-953.

Schultz, V.(2015). Taking sex discrimination seriously. Denver University Law Review, 91(5), 995-1119.

Firth, M.(1982). Sex discrimination in job opportunities for women. Sex Roles, 8(8), 891-901.

Chant, S.H.(2003). Female household headship and the feminisation of poverty: facts, fictions and forward strategies [online]. London: LSE Research Online. January, 2006. http://eprints.lse.ac.uk/archive/00000574

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

Johnson, R. W., & Neumark, D. (1996). Age discrimination, job separations, and employment status of order workers: evidence from self-reports. The Journal of Human Resources, 32(4), 779-811.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: racial discrimination in employment, housing, credit, and consumer markets. Annual Review of Sociology, 34(1), 181-209.

Gammarano, R. (2020). Education pays off, but you have to be patient. International Labour Organization. August 18, 2020. https://ilostat.ilo.org/education-pays-off-but-you-have-to-be-patient/

Stuart, E. A., & Rubin, D. B., & Osborne, J. (2007). Matching methods for causal inference: Designing observational studies. ResearchGate. February, 2007. https://www.researchgate.net/profile/Donald_Rubin3/publication/228519896_Matching_methods_for_causal_inference_Designing_observational_studies/links/0c96051a4e77565396000000.pdf

Rubin, D. B. (1973). Matching to remove bias in observational study. Biometrics, 29(1), 159-183.

Lee, B. K., & Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Stat Med., 29(3), 337–346

Wolla, S. A., & Sullivan, J. (2017). Education, income, and wealth. Federal reserve bank of St. Louis. January, 2017. https://research.stlouisfed.org/publications/page1-econ/2017/01/03/education-income-and-wealth/

Alexander, R. (2020). Difference in differences. Telling stories with data. Nov. 5, 2020. https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html

Prabhakaran, S. (2016). Top 50 ggplot2 visualizations - the master list. r-statistics.co. 2016-17. http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Wickham et al. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Miller, E. (2020). haven: Import and export 'spss','stata' and 'sas' files. Import foreign statistical formats into R via the embedded 'ReadStat' C library. http://haven.tidyverse.org, https://github.com/tidyverse/haven, https://github.com/WizardMac/ReadStat

Jennifer, B. (NA). gapminder: Data from Gapminder. https://github.com/jennybc/gapminder, http://www.gapminder.org/data/, https://doi.org/10.5281/zenodo.594018.

Robinson, D., Hayes, A., & Couch, S. (2020). broom: Convert statistical objects into tidy tibbles. https://broom.tidymodels.org/, https://github.com/tidymodels/broom

Wickham, H., François, R., Henry, L., & Müller, K. (2020). dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Gelman, A., & Su, Y. S. (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm

Fox, J., & Weisberg, S. (2019). An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Arnold et al. (2019). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. http://github.com/jrnold/ggthemes

Rich, B. (2020). table1: Tables of Descriptive Statistics in HTML. https://github.com/benjaminrich/table1

Wickham, H., & Seidel, D. (2020). scales: Scale Functions for Visualization. https://scales.r-lib.org, https://github.com/r-lib/scales.